# POWER-LAW-LIKE DISTRIBUTIONS: A PRACTICAL SURVEY

A. CARRIE,* Department of Economics,
University of Dublin, Trinity College, Dublin, Ireland

## ABSTRACT

Unrelated datasets from biology, economics, computer science, and many other disciplines follow power law distributions, characterized by a straight line in the log-log rank-frequency plot. This universality, along with the tempting prospect of a common underlying generative process, has attracted significant research interest. Upon closer inspection, many of these datasets show slight or pronounced curvature. In light of this, several alternative distributions have been proposed in the literature. The lure of the power law, however, is extremely strong, and these alternatives are rarely fitted. This paper reviews these alternative distributions and fits them to a standardized collection of power law datasets. The practicalities of fitting these distributions are discussed. The hope is that presenting these distributions in a user-friendly and systematic format and testing them against some canonical datasets will facilitate their use within the power law literature.

**Keywords:** Power law distribution, DGX, lognormal distribution, maximum likelihood estimation, rank-frequency plot

## INTRODUCTION

Many datasets that are described as following a power law (i.e., having a linear probability distribution in log-log coordinates) do, in fact, show some inconvenient curvature. There are several alternative distributions that allow for such curvature in the literature; however, it is not straightforward for researchers to utilize these because they do not share consistent notation or statistical methodology. Ideally, we should have access to a readily available toolkit of skew distributions that are straightforward to use and interpret and that can be easily compared with each other. In this paper, I undertake an informal survey of some of these skew distributions that I hope may contribute to the development of such a toolkit. This work is at an early stage, so references should be made to other sources before using any result included here. I first define the terminology and notation used throughout this paper. I then describe the statistical distributions, their functional forms, and how their parameters may be estimated from sample data. I work through the Beowulf dataset in detail as an example. This is followed by summary results for several datasets and finally by concluding remarks.

* *Corresponding author address*: Ana Carrie, Department of Economics, Trinity College, Dublin 2, Ireland; email: carriea@tcd.ie.

## NOTATION AND TERMINOLOGY

Observations $x_1,\ldots,x_n$ are sorted from largest to smallest so that the subscript $i$ corresponds to the rank of the observation, the largest observation being ranked first and the smallest $n$'th. The value of an observation is known as a frequency. The origin of this unfortunate convention is that in many cases, particularly in the early terminology-formation years, the values being studied were frequencies of occurrence. For instance, our first sample dataset will be one studied by Zipf: the frequency of words appearing in the text of Beowulf. Our vector of observations is also converted into the form $f_1,\ldots,f_m$ (frequency) and $c_1,\ldots,c_m$ (count), where $c_i$ represents the count of observations equal to $f_i$, such that the sample size $n = \sum_{i=1}^{m} c_i$. That is, if our original vector of observations is (30, 5, 3, 2, 2, 1, 1, 1), then we will have a frequency vector (30, 5, 3, 2, 1), with each value appearing once, and a count vector (1, 1, 1, 2, 3). Summing the values in count (1 + 1 + 1 + 2 + 3 = 7) tells the total sample size. The use of "count" and "frequency" in this very specific manner is problematical, since these words are interchangeable in everyday speech. If you are familiar with using the function count ( ) in a summary calculation, this should help keep you oriented correctly. In general, it is advisable to read any use of this terminology very carefully to make sure of the author's intended meaning.[1]

## THE DISTRIBUTIONS

### Power Law

The probability density function (PDF) of the power law distribution is given by:

$$f(x) \propto x^{-(k+1)}, \tag{1}$$

where the parameter $k$ determines the steepness of the slope. The probability mass function in the upper tail (PMUF) is

$$\overline{F}(x) \propto x^{-k}. \tag{2}$$

Taking logarithms of $y = x^{-k}$ gives $\log y = -k \log x$, which illustrates the trademark power law linear relationship in log-log coordinates (Adamic 2005). In principle, one could determine the distribution parameter $k$ either by fitting a straight line to PDF data as given by Equation 1 or to PMUF data as given by Equation 2. In practice, the PDF does not yield reliable results, so the PMUF is used.

The power law can also be fit by using rank-frequency data. In a log-log rank-frequency plot, the parameter $k$ is derived from the slope $-b$ by $k = 1/b$. In the traditional Zipf distribution, both $b$ and $k$ are equal to 1 (Adamic 2005). Note that there is no intercept term in the PMUF regression, since the line fitted must pass through (1,1), which is (0,0) in the log-log plot. For rank-frequency data, we fit a straight line with an arbitrary intercept, the fitted intercept giving us the scale of the object with rank 1.

---

[1] I strongly considered referring to frequency data by another name. However, the term "rank-frequency plot" is widely used, so it was preferable to explain and use the terminology.

Unfortunately, all of the above apply only to continuous data. Our data are discrete; in fact, in the Beowulf example and many others, the data take only integer values. Following the example of Bi et al. (2001), we can derive the discrete probability function (point-mass function) as follows:

$$p(x) = \frac{x^{-(k+1)}}{\sum_{x=1}^{\infty} x^{-(k+1)}} \tag{3}$$

or

$$p(x) = \frac{x^{-(k+1)}}{\varsigma(k+1)} , \tag{4}$$

where $\zeta(n) = \sum_{i=1}^{\infty} \frac{1}{i^n}$ is the Riemann zeta function (Weisstein 2005). We will use maximum likelihood to determine the distribution parameter *k*. When independent, identically distributed data are assumed, the likelihood function is

$$L(k) = \prod_{i=1}^{n} P(x_i) = \prod_{i=1}^{n} \frac{x_i^{-(k+1)}}{\zeta(k+1)} . \tag{5}$$

The log likelihood function has a simpler form:

$$l(k) = -(k+1) \sum_{i=1}^{n} \log x_i - n \log\left[\zeta(k+1)\right] , \tag{6}$$

or in terms of count-frequency data:

$$l(k) = -(k+1) \sum_{i=1}^{m} c_i \log f_i - \left( \sum_{i=1}^{m} c_i \right) \log\left[\zeta(k+1)\right] . \tag{7}$$

## Lognormal/DGX

The most well-known alternative distribution is the lognormal. While the power law has a straight line in log-log coordinates, the lognormal is parabolic. The PDF is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}x} \exp\left[ -\frac{(\ln x - \mu)^2}{2\sigma^2} \right] . \tag{8}$$

The discretized form of the lognormal distribution, known as the discrete Gaussian exponential (DGX) (Bi et al. 2001), has this PDF:

$$p(x) = \frac{A(\mu,\sigma)}{x} \exp\left[ -\frac{(\ln x - \mu)^2}{2\sigma^2} \right], \tag{9}$$

where $A$ is a normalization constant given by:

$$A(\mu,\sigma)=\left\{\sum_{j=1}^{\infty}\frac{1}{j}\exp\left[-\frac{(\ln j-\mu)^2}{2\sigma^2}\right]\right\}^{-1}. \tag{10}$$

When independent, identically distributed data are assumed, the log likelihood function is

$$l(\mu,\sigma)=n\ln A(\mu,\sigma)-\sum_{i=1}^{n}\left[\ln x_i+\frac{(\ln x_i-\mu)^2}{2\sigma^2}\right]. \tag{11}$$

By substituting count and frequency data, the log likelihood expression becomes

$$l(\mu,\sigma)=\left(\sum_{i=1}^{m}c_i\right)\ln A(\mu,\sigma)-\sum_{i=1}^{m}c_i\left[\ln f_i+\frac{(\ln f_i-\mu)^2}{2\sigma^2}\right]. \tag{12}$$

The DGX reduces to the power law as $\mu\to-\infty$.

## Stretched Exponential

The stretched exponential distribution is a generalization of the exponential distribution. The PDF is defined as

$$f(x)=c\left(\frac{x^{c-1}}{x_0^c}\right)\exp\left[-\left(\frac{x}{x_0}\right)^c\right], \tag{13}$$

with the cumulative distribution function (CDF) being

$$F(x)=P(X\le x)=\exp\left[-\left(\frac{x}{x_0}\right)^c\right] \tag{14}$$

for $c \le 1$. When $c = 1$, this reduces to the exponential distribution (Laherrère and Sornette 1998).

The stretched exponential produces a straight line when the natural logarithm of the rank is plotted against observed values raised to the power $c$:

$$x_i^c=-a\ln i+b. \tag{15}$$

The three parameters of the distribution are $a$, $b$, and $c$, with $x_0=a^{\frac{1}{c}}$.

The authors provide no algorithm for fitting the stretched exponential. Thus far, the simplest method I have found is one of brute force. Allow $c$ to take each of the values in (0.001, 0.002, …, 0.999, 1.000), or the required search precision, and proceed to fit the linear model specified in Equation 15 to the vector of observations $x_1^c,\ldots,x_n^c$. Choose the value of $c$ that corresponds to the highest regression $R^2$, and $a$ and $b$ are then obtained from the corresponding linear model.

### Parabolic Fractal

The parabolic fractal is another second-order polynomial extension of the linear power law, but while the lognormal is a parabola in log-log frequency-count, the parabolic fractal is a parabola in log-log rank-frequency:

$$\log x_i = \log x_1 - a \log i - b(\log i)^2 . \tag{16}$$

When $b = 0$, this reduces to the power law. Since a concave parabola has a maximum value, the theoretical maximum observation (regardless of sample size) can be calculated as follows:

$$x_{max} = x_1 e^{\left(\frac{a^2}{4^b}\right)} . \tag{17}$$

The parabolic fractal can be fit by using linear regression on log $i$ and $(\log i)^2$.

A future task is to develop discretized versions of both the stretched exponential and the parabolic fractal so that they can be directly compared with the discrete power law and DGX.

### Other Distributions

This is not an exhaustive list, and new distributions are being developed all the time, such as the double Pareto, which has two straight-line segments connected at a transition point (Mitzenmacher 2003) rather than a single straight line as in the standard Pareto/power law.

## SAMPLE DATASET

Beowulf, one of the earliest surviving poems in English, was a source text for Zipf's study of the frequency with which words appear in the written language (Zipf 1965). The text of Beowulf was obtained from Project Gutenberg, and a word count list (concordance) was prepared, the start of which is shown in Table 1.

Table 2 has word frequencies in the first column and the number/count of words that appear with said frequency in the second column. There are 1,611 words that appear only once in the text, and the most common word (THE) appears 1,587 times. Rank is also given for the highest-ranked observations:

**TABLE 1** Concordance of Beowulf

| Frequency | Word |
|---|---|
| 1 | ABANDONED |
| 1 | ABEL |
| 2 | ABIDE |
| 1 | ABJECT |
| 3 | ABLE |
| 4 | ABODE |
| 6 | ABOUT |
| 2 | ABOVE |
| 1 | ABROAD |
| 2 | ACCURSED |

**TABLE 2** Frequency, count, and rank data for Beowulf

| Frequency | Count | Rank | Word |
|---|---|---|---|
| 1 | 1,611 | | |
| 2 | 548 | | |
| 3 | 293 | | |
| 4 | 180 | | |
| 5 | 115 | | |
| 6 | 93 | | |
| 7 | 61 | | |
| 8 | 49 | | |
| … | … | | |
| 163 | 1 | 8 | HIM |
| 222 | 1 | 7 | FOR |
| 229 | 1 | 6 | WAS |
| 276 | 1 | 5 | WITH |
| 321 | 1 | 4 | THAT |
| 408 | 1 | 3 | HIS |
| 636 | 1 | 2 | AND |
| 1587 | 1 | 1 | THE |

A natural first step in our exploration is to graph the count-frequency data in both linear and logarithmic scale (Figure 1). The log-log graph on the right suggests a linear relationship, albeit with rather messy data for high-frequency words. Figure 2 shows that this plot is not suitable for curve fitting. Although the values obtained for *k* will not be correct for our discrete data, by way of illustration, Figure 3 shows the CDF and rank-frequency plots.

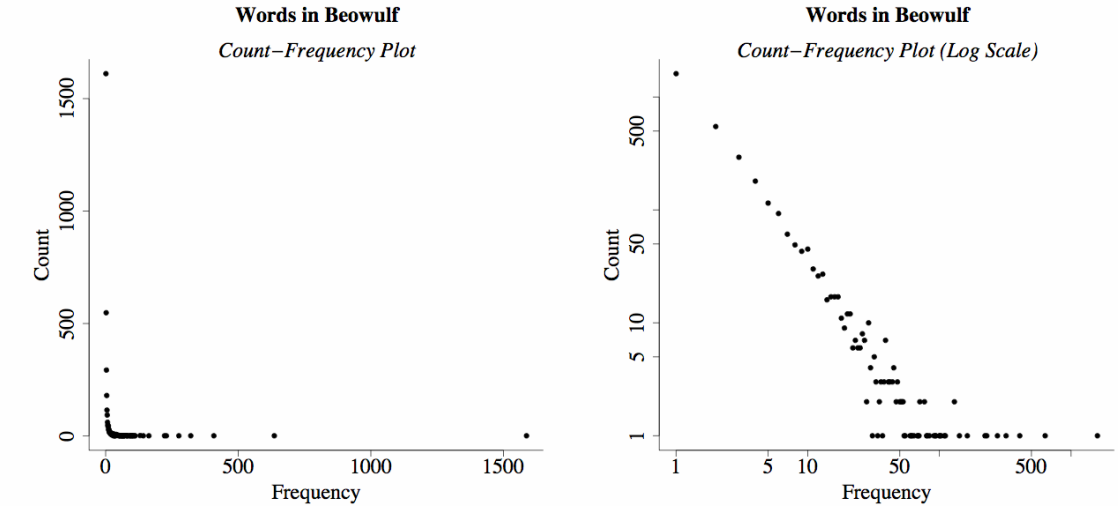


**FIGURE 1** Count and frequency data in linear and logarithmic scale

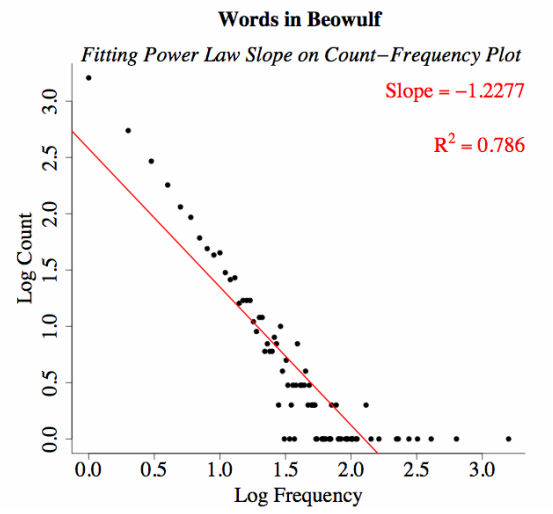


**FIGURE 2** Fitting power law slope on count-frequency plot

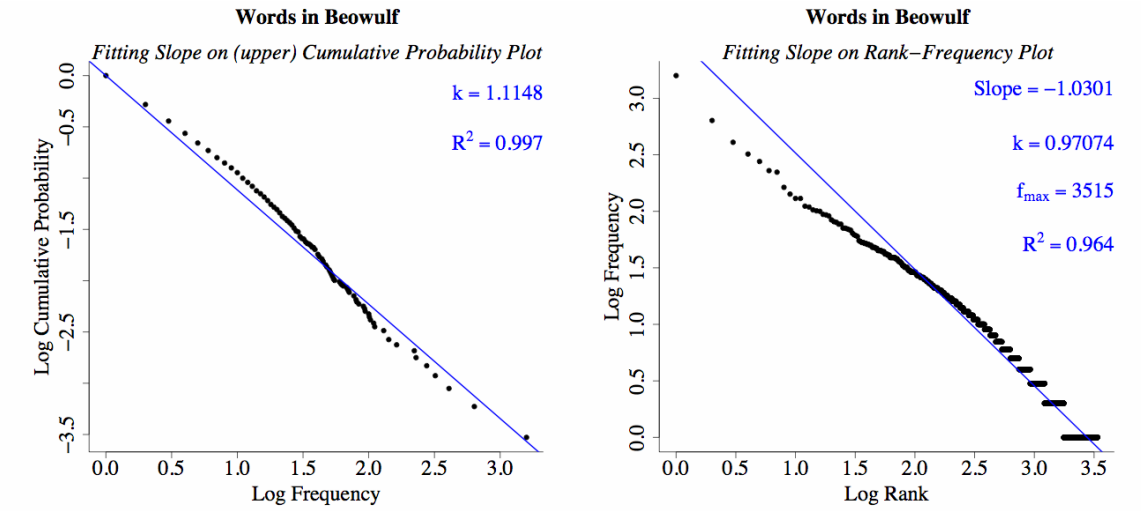


**FIGURE 3** CDF and rank-frequency plots

We can see some evidence of curvature in the CDF plot, and even more in the rank-frequency plot. We will now calculate the power law distribution parameter $k$ and the DGX distribution parameters $\mu$ and $\sigma$ by maximizing the respective log likelihood functions. To compare these two distributions, we can define an error statistic (denoted *ERR*), which is a straightforward extension of the mean squared error (MSE):

$$ERR = \sum_{i=1}^{m} \frac{[nP(f_i) - c_i]^2}{m}. \tag{18}$$

We see that the DGX has a much lower error statistic; it is 12, compared with 358 for the power law. Since the DGX is, in effect, a generalization of the power law, this is to be expected.

Moving on to the stretched exponential and parabolic fractal (Figure 4), we will be able to compare them with each other, but not, for now, with the discrete power law and DGX (Figure 5). We can define an error statistic for rank-frequency data by

$$ERR = \sum_{i=1}^{n} \frac{[F(i) - x_i]^2}{n}, \tag{19}$$

where $F(i)$ is the predicted frequency for the observation of rank $i$.

For this dataset, the stretched exponential and parabolic fractal give similarly shaped fitted curves and similar error statistics. We see that both curves miss the handful of highest-ranked observations by a considerable amount. This may be a feature of the rank-frequency plot, which has n data points and thus places more emphasis on common, small events; in the PDF plot, these small events are aggregated together.

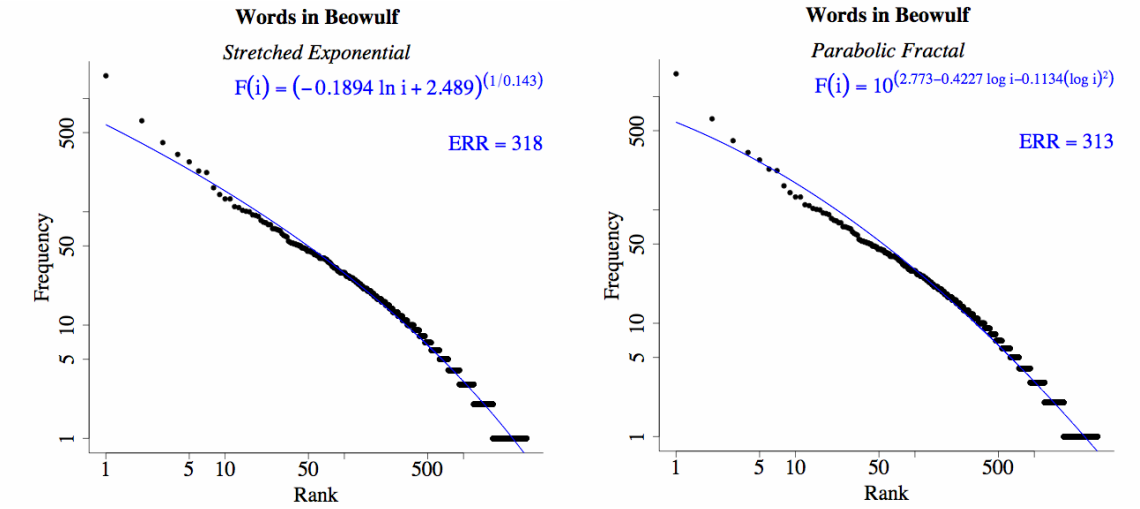


**FIGURE 4** Stretched exponential and parabolic fractal

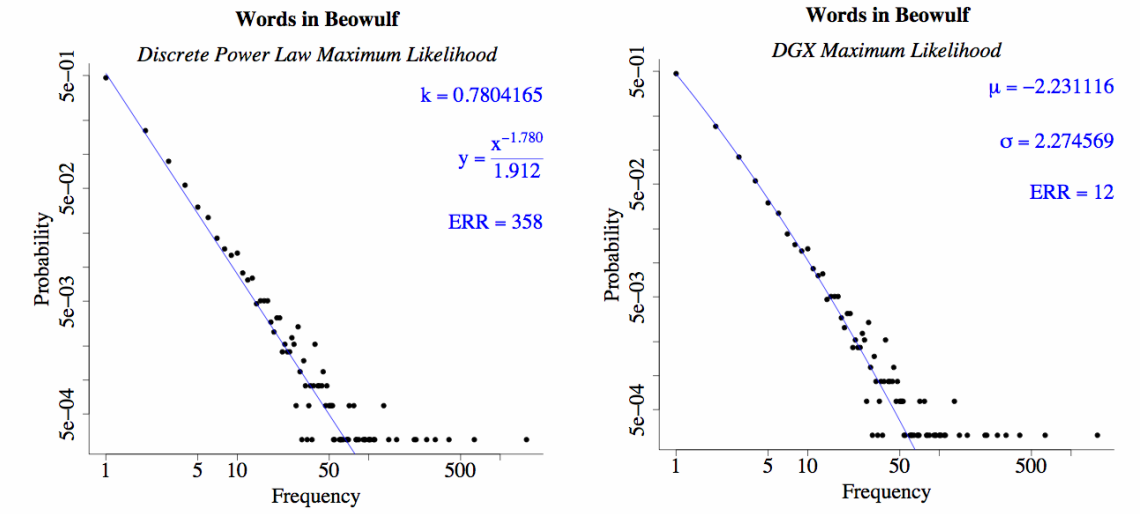


**FIGURE 5** Discrete power law and DGX

## OTHER DATASETS

### Genera and Species of Snake

The number of species in a genus, for a family of plants or animals, has a skew distribution. I present two datasets here. One is from Yule (1925), which was quoted from an earlier work by Willis, which collated the data from the *Catalogue of the Snakes in the British Museum* by G.A. Boulenger, published in 1893 (Figure 6). The other is an updated version with 2005 data (Uetz and Heidelberg 2005) (Figure 7). There are 293 genera and 1,475 species in the 1893 dataset, and 463 genera and 3,002 species in the 2005 dataset.

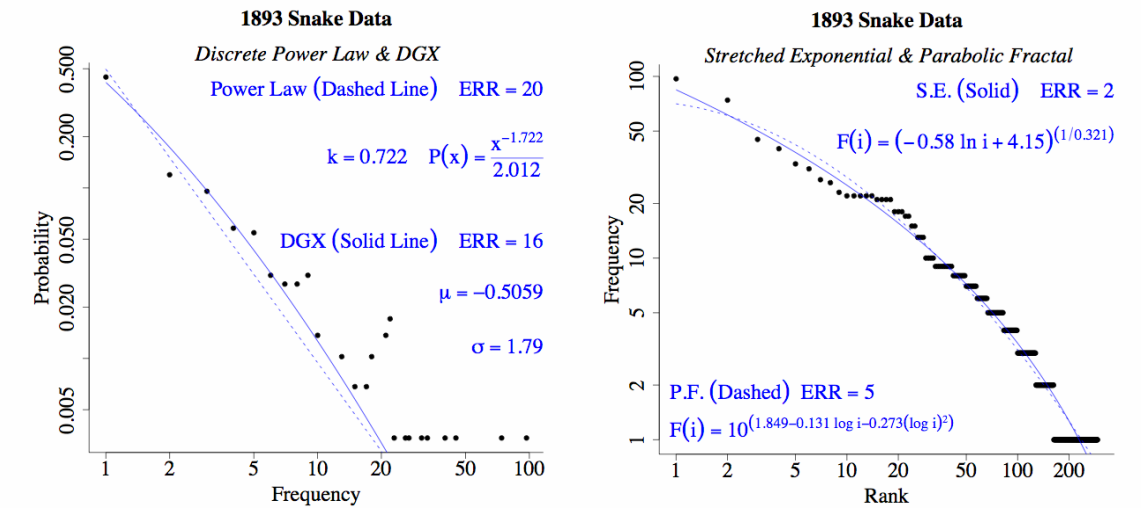


**FIGURE 6** Year 1893 snake data

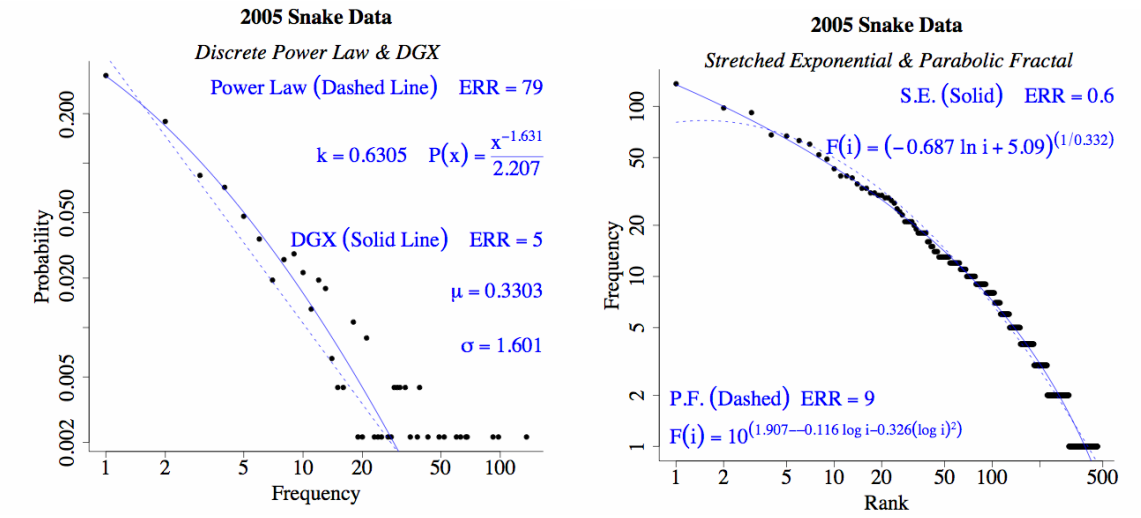


**FIGURE 7** Year 2005 snake data

For both 2005 and 1893 data, we see that the DGX has a lower error statistic, as we expect. The stretched exponential is also a better fit in both cases, and in the 2005 data, it seems to match even the largest events. The parabolic fractal, in addition to having a poor fit, also has a positive coefficient for log(*i*) in the 2005 data, which violates its specification.

### U.S. Cities

Here we look at the distributions of population in U.S. cities with more than 100,000 people. The DGX again outperforms the power law. The parabolic fractal in this instance has a slightly lower error statistic than the stretched exponential, but this is probably not a significant difference (Figure 8).

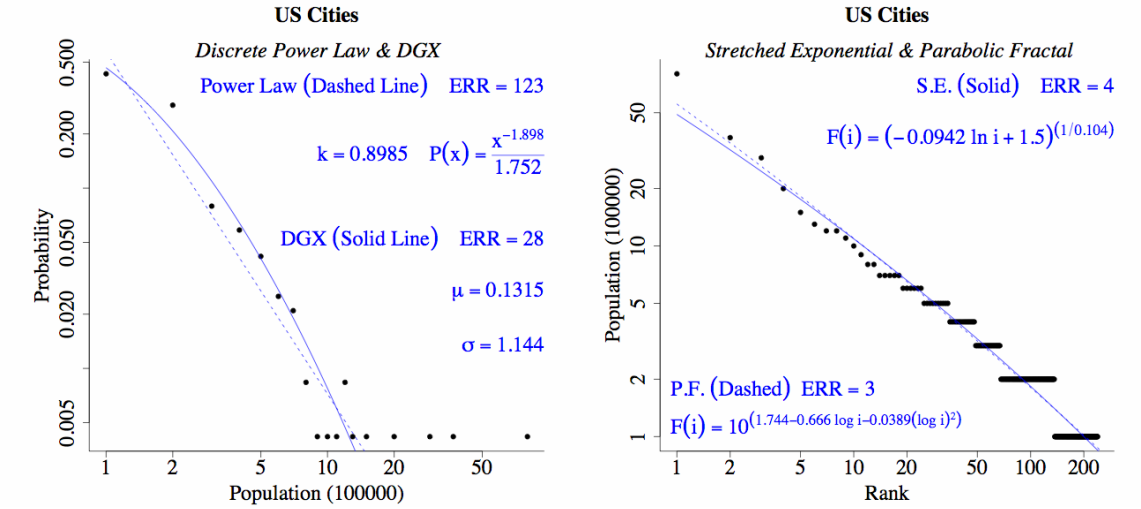


**FIGURE 8** U.S. city sizes

## Isle of Man Surnames

The DGX has an unfortunate quirk that means that for certain data, the normalization constant takes a very long time to converge. This means that the computations can be prohibitively time-consuming. Hence, for data pertaining to the distribution surnames of families living in the Isle of Man, we show only the stretched exponential and parabolic fractal (Figure 9). Again, the parabolic fractal has a negative coefficient for one of its terms, which is not valid according to its definition. Visually, the stretched exponential seems to fit both extremes, but it misses the curvature in the middle of this dataset.

## DISCUSSION AND CONCLUSIONS

With so much academic interest in power laws, much more research is needed on the probability distributions that describe skew data, including the development of standardized criteria for discriminating between alternative distributions. Many of the conventional tools are based on a distribution having a finite mean or following a Gaussian error distribution; thus, they are not helpful for dealing with skew data. Because this research interest is interdisciplinary, consistent terminology and notation are all the more crucial. Distributions tend to be invented in response to a particular research problem and so have "baggage" from the academic or industrial realm in which they arose.

The DGX and discrete power law were fit by using maximum likelihood, and their distribution functions explicitly are acknowledged the discrete nature of the data. The stretched exponential and parabolic fractal were fit by using linear regression, implicitly assuming continuous data, and they were fit in the rank-frequency plot rather than a frequency-count or frequency-PDF plot. It is not clear at this point whether these two approaches will turn out to be complementary, each highlighting different and useful aspects of the data, or whether one will emerge to be "correct."

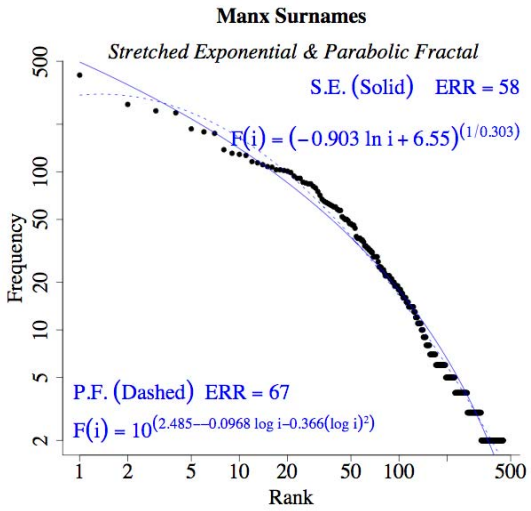


**FIGURE 9** Distribution of surnames in the Isle of Man

There may or may not be a "best" alternative distribution that is a better fit for all or nearly all datasets. For the datasets considered here, the DGX was a better fit than the discrete power law, which was expected, since the DGX is a generalization of the power law. The parabolic fractal proved problematic, since it should be strictly decreasing, but for several datasets, the fit produced by linear regression led to negative values for the "a" coefficient. The stretched exponential did not have this difficulty, and it had a better or comparable error statistic to the parabolic fractal.

I am looking forward to continuing this work and incorporating additional distributions and datasets. There are plenty of practical and theoretical challenges involved in working with skew distributions, and the development of a statistical methodology will be a vital component of research in the years to come.

## REFERENCES

Adamic, L.A., 2005, *Zipf, Power-laws, and Pareto — A Ranking Tutorial;* available at http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html. Accessed May 26, 2005.

Bi, Z., et al., 2001, *The "DGX" Distribution for Mining Massive, Skewed Data,* San Francisco, CA.

Laherrère, J., and D. Sornette, 1998, "Stretched Exponential Distributions in Nature and Economy: 'Fat Tails" with Characteristic Scales," *The European Physical Journal B* 2, 525–539.

Mitzenmacher, M., 2003, "A Brief History of Generative Models for Power Law and Lognormal Distributions," *Internet Mathematics* 1, 226–251.

Uetz, P., and EMBL Heidelberg, 2005, *The EMBL Reptile Database;* available at http://www.reptile-database.org/. Accessed Sept. 19, 2005.

Weisstein, E.W., et al., 2005, "Riemann Zeta Function," in *Mathworld — A Wolfram Web Resource;* available at http://mathworld.wolfram.com/RiemannZetaFunction.html.

Yule, G.U., 1925, "A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J.C. Willis, F.R.S.," *Philosophical Transactions of the Royal Society of London, Series B* 213, 21–87.

Zipf, G.K., 1965, *Human Behaviour and the Principle of Least Effort,* New York: Hafner Publishing Co.